

## STATISTICS 1 – POSSIBLE THEORETICAL QUESTIONS

- ✓ Process of designing a mathematical model
  1. A real world problem is observed
  2. A mathematical model is devised
  3. The model is used to make predictions
  4. Experimental data is collected
  5. Predicted and observed outcomes are compared
  6. Statistical tests are used to assess how well the model describes the problem
  7. The mathematical model is refined if necessary
- ✓ Advantages
  1. Quick and easy to produce
  2. A simplification of a more complex situation
  3. Can help improve understanding
  4. Enables prediction to be made
  5. Can help provide control
- ✓ Disadvantages
  1. It does not include all aspects of the problem
  2. It may only work in certain situations
- ✓ The median is preferred when the data is skewed, as it is not easily affected by extreme values
- ✓ The mean is used when the data is (roughly) symmetrical, as it uses all of the data.
- ✓ If asked to compare and contrast (usually with a box plot), you need to make three statements
  - Location (median)
  - Dispersion (IQR or range)
  - Skewness (positive, negative or symmetric)
- ✓ Coding

If  $y = \frac{x-a}{b}$ , where  $x$  is the original and  $y$  the coded data then  $\bar{y} = \frac{\bar{x}-a}{b}$  and  $\sigma_y = \frac{\sigma_x}{b}$

- ✓ Histograms are used when the data is continuous
- ✓ The basic feature of a histogram is that the area is proportional to the frequency
- ✓ If events  $A$  and  $B$  are statistically independent then  $P(A \cap B) = P(A) \times P(B)$
- ✓ If events  $A$  and  $B$  are mutually exclusive then  $P(A \cup B) = P(A) + P(B)$  or  $P(A \cap B) = 0$  (on a Venn diagram this would be shown as two distinct circles)
- ✓ Conditional probability questions can be solved either using the formula or by considering the “shrunk” sample space approach.
- ✓ A linear regression model may be used when the points on the scatter plot lie roughly along a straight line or when the product moment correlation coefficient is close to 1 or -1
- ✓ Estimates obtained through extrapolation should always be viewed with caution
- ✓ The explanatory or independent variable is the one that is plotted on the  $x$ -axis and is set independently of the response or dependent variable which is plotted on the  $y$ -axis. E.g. in a study of how height varies with age, age is the independent and height the dependent variable (it is because the age changes that height changes and not the other way round!)
- ✓ **Always give your answers correct to 3 significant figures BUT do make sure that you use the full accuracy of any intermediate answers in order to get the right final answer**
- ✓ Wherever possible, give your answers in context to the question (especially when you are asked to interpret)